UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

MIKE HUCKABEE, RELEVATE GROUP, DAVID KINNAMAN, TSH OXENREIDER, LYSA TERKEURST, and JOHN BLASE *on behalf of themselves and all others similarly situated,*

Plaintiffs,

v.

META PLATFORMS, INC., BLOOMBERG L.P., BLOOMBERG FINANCE, L.P., MICROSOFT CORPORATION, and THE ELEUTHERAI INSTITUTE,

Defendants.

Case No.

CLASS ACTION

JURY TRIAL DEMANDED

CLASS ACTION COMPLAINT

Plaintiffs Mike Huckabee, Relevate Group, David Kinnaman, Tsh Oxenreider, Lysa TerKeurst, and John Blase ("Plaintiffs") bring this Class Action Complaint against Defendants Meta Platforms, Inc., Bloomberg L.P., Bloomberg Finance, L.P., Microsoft Corporation, and The Eleutherai Institute ("Defendants"), and allege as follows:

NATURE OF THE ACTION

1. Large language models ("LLMs") are a type of artificial intelligence ("AI") that are designed to understand and generate human language. These models are characterized by their size and complexity, with millions of parameters that allow them to process and generate text in a way that appears highly intelligent and contextually relevant.

2. Developing LLMs, however, requires a massive amount of text data. Internet sources, such as Wikipedia, provide some training text from which LLMs can be developed, but higher-quality input is required to develop humanlike answers to language prompts.

3. Because they are a substantial source of written language, books are often used in libraries of information (or datasets) to create more sophisticated LLMs. While using books as part of datasets is not inherently problematic, using pirated (or stolen) books does not fairly compensate authors and publishers for their work.

4. Defendant The EleutherAI Institute ("EleutherAI") has trained and released several series of LLMs and the codebases used to train them. Touting that these LLMs were the "largest or most capable LLMs available at the time and have been widely used since in open-source research applications,"[1] EleutherAI recognizes that large datasets of sophisticated information are necessary to develop humanlike responses to prompts.

5. EleutherAI's dataset, called "The Pile," contains 800 gigabytes of diverse text for language modeling. The Pile was introduced via an open-source paper in December 2020.[2]

6. EleutherAI touts the Pile's "diversity in data sources," which purportedly "improves general cross-domain knowledge." According to EleutherAI, "models trained on the Pile show moderate improvements in traditional language modeling benchmarks[.]"[3]

7. As part of that "diversity in data sources," the Pile includes "Books3," a dataset of information scraped from a large collection of approximately 183,000 pirated ebooks, most of which were published in the past 20 years.[4]

[1] EleutherAI, Research, *available at* https://www.eleuther.ai/research (last accessed Oct. 6, 2023).

[2] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, *available at* https://browse.arxiv.org/pdf/2101.00027.pdf (last accessed Oct. 6, 2023).

[3] *Id.*

[4] The EleutherAI paper cites to a Twitter (now X) thread from Shawn Presser, announcing that nearly "200k plaintext books" were available via plaintext for processing by LLMs. He includes a citation to "an interesting [Digital Millennium Copyright Act] policy. I urge you to read it." *See* Shawn Presser, Oct. 25, 2020 Tweet, *available at* https://twitter.com/theshawwn/status/1320282154488766464. The link is to a profane performance and is, likely, a joke acknowledging the copyright violation.

8. The EleutherAI paper that introduced the Pile reveals that the Books3 dataset comprises 108 gigabytes of data, or approximately 12% of the dataset, making it the third largest component of The Pile by size.[5] The EleutherAI paper also describes the contents of Books3:

> Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker … Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset (BookCorpus2). We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.[6]

9. The Pile, and Books3 specifically, quickly became a popular training data set for some of the biggest, most prolific, and most successful companies developing AI technology, including Defendants Microsoft Corporation ("Microsoft"), Meta Platforms, Inc. ("Meta"), Bloomberg L.P., and Bloomberg Finance L.P. (together with Bloomberg L.P., "Bloomberg").

10. Microsoft, Meta, and Bloomberg each used Books3 to train LLMs, with the full knowledge and understanding that the datasets they were using to train their LLMs were assembled from copyrighted works, including copyrighted works of the Plaintiffs and Members of the Class.

11. None of the Defendants sought or obtained licenses to use the copyrighted works from Books3, and the Defendants knew that the parties responsible for assembling Books3 were not licensed, or otherwise legally permitted, to disseminate those works.

12. Microsoft, Meta, and Bloomberg chose to train their LLMs using pirated and stolen works for the purpose of making a profit. Accordingly, Plaintiffs and the Class were injured, and are entitled to damages.

[5] See Github, "The Pile Replication Code" (available at https://github.com/EleutherAI/the-pile), breaking down by raw size and weight the various subsets of data in The Pile.

[6] *Id.* at 3-4.

JURISDICTION AND VENUE

13. This Court has subject matter jurisdiction under 28 U.S.C. § 1331 because this case partly arises under the Copyright Act, 17 U.S.C. § 501, and the Digital Millennium Copyright Act, 17 U.S.C. § 1202.

14. This Court has supplemental subject matter jurisdiction over Plaintiffs' state law claims, under 28 U.S.C. § 1367, because Plaintiffs' state law claims arise from the same common nucleus of operative facts as Plaintiffs' claims under federal law.

15. Jurisdiction is appropriate as to Defendants Bloomberg, L.P. and Bloomberg Finance, L.P. because they are headquartered and have their principal place of business in the State of New York and in this district.

16. Jurisdiction is appropriate as to Defendant EleutherAI because, upon information and belief, it maintains such continuous and systematic contacts with New York and this district that it is essentially "at home" here.[7] Alternatively, as alleged in more detail below, EleutherAI has purposefully availed itself of the privilege of conducting activities within the State of New York in a series of actions that are directly related to the transactions and causes of action alleged in this suit. Upon information and belief, it transacted with Defendants and others in New York City to provide them access to the Books3 dataset containing Plaintiffs' stolen, copyrighted works. In addition, EleutherAI maintains contracts and agreements which specifically subject it to jurisdiction and venue in the state of New York.

17. Jurisdiction is appropriate as to Defendants Meta and Microsoft because each of those Defendants purposefully availed themselves of the privilege of conducting activities within

[7] Davide Castelvecchi, "Open-Source AI Chatbots Are Booming—What Does This Mean for Researchers?," 618 NATURE 891, 891 (June 29, 2023) (describing EleutherAI as "an AI research institute in New York City").

the State of New York, and these contacts were not random, isolated, or fortuitous. Each Defendant deliberately reached out beyond its home state to obtain and access the Books3 dataset from EleutherAI, "an AI research institute in New York City."[8] Such access allowed each Defendant to infringe on the copyrights of Plaintiffs and the Class. Because each Defendant's purposeful transactions with EleutherAI in New York were part and parcel of its violation of Plaintiffs' copyrights, Plaintiffs' claims arise out of each Defendants' contacts with this forum, such that this Court may exercise jurisdiction over them for the claims alleged herein.

18. Venue is proper in this district under 28 U.S. Code § 1391 because at least one of the named Defendants is a corporate resident of the State of New York and of this district.

PARTIES

19. **Plaintiff Mike Huckabee.** Plaintiff Mike Huckabee ("Huckabee") is the former Governor of Arkansas and a former candidate for President of the United States. He is a citizen and resident of Pulaski County, Arkansas. Plaintiff Huckabee authored several works, including the following:

a. *A Simple Government: Twelve Things We Really Need from Washington (and a Trillion That We Don't)* (bearing copyright Registration Number TX0007354200, registered on April 14, 2011);

b. *God, Guns, Grits, and Gravy* (bearing copyright Registration Number TX0007999502, registered on February 2, 2015); and

c. *Rare, Medium, or Done Well: Make the Most of Your Life* (bearing copyright Registration Number TX0008710363, registered on January 30, 2019).

[8] *Id.*

20. **Plaintiff Relevate Group, Inc.** Plaintiff The Relevate Group, Inc. ("Relevate") is a Georgia non-profit corporation that is licensed to do business in Tennessee and has its principal place of business in Williamson County, Tennessee. David Kinnaman and Gabriel Lyons co-authored the work *unChristian: What a New Generation Really Things about Christianity… and Why It Matters*. Mr. Lyons assigned his copyright to The Relevate Group, Inc., d/b/a Fermi Project (bearing copyright Registration Number TX0007009963, registered on April 25, 2008).

21. **Plaintiff David Kinnaman.** Plaintiff David Kinnaman ("Kinnaman") is a citizen and resident of Tarrant County, Texas. Plaintiff Kinnaman co-authored the above-mentioned work *Unchristian: What a New Generation Really Things about Christianity… and Why It Matters*, and holds the copyright in his personal name (bearing copyright Registration Number TX0007009963, registered on April 25, 2008).

22. **Plaintiff Tsh Oxenreider.** Plaintiff Tsh Oxenreider ("Oxenreider") is a citizen and resident of Austin County, Texas. Plaintiff Oxenreider authored several works, including the following:

a. *At Home in the World: Reflections on Belonging While Wandering the Globe* (bearing copyright Registration Number TX0008407149, filed on April 26, 2017);

b. *Notes from a Blue Bike: The Art of Living Intentionally in a Chaotic World* (bearing copyright Registration Number TX0007879266, filed on February 18, 2014); and,

c. *Organized Simplicity: The Clutter-Free Approach to Intentional Living* (bearing copyright Registration Number TX0007290539, registered on November 12, 2010).

23. **Plaintiff Lysa TerKeurst.** Plaintiff Lysa TerKeurst ("TerKeust") is a citizen and resident of St. Johns County, Florida. Plaintiff TerKeust authored several works, including the following:

a. The No. 1 *New York Times* bestseller *Uninvited* (bearing copyright Registration Number TX0008318547, registered on August 17, 2016);

b. *Capture His Heart: Becoming the Godly Wife Your Husband Desires* (bearing copyright Registration Number TX0005618018, registered on May 2, 2002);

c. *Embraced: 100 Devotions to Know God is Holding You Close* (bearing copyright Registration Number TX0008564815, registered on April 4, 2018);

d. *It's Not Supposed to Be This Way Study Guide: Finding Unexpected Strength When Disappointments Leave You Shattered* (bearing copyright Registration Number TX0008677793, registered on November 20, 2018);

e. *Uninvited - Study Guide* (bearing copyright Registration Number TX0008318550, registered on August 17, 2016);

f. *Made to Crave Devotional: 60 Days to Craving God, Not Food* (bearing copyright Registration Number TX0007461682, filed on December 8, 2011);

g. *The Best Yes: Making Wise Decisions in the Midst of Endless Demands* (bearing copyright Registration Number TX0007939136, registered on August 20, 2014); and

h. *Unglued: Making Wise Choices in the Midst of Raw Emotions* ((bearing copyright Registration Number TX0007580702, registered on August 9, 2012).

24. **Plaintiff John Blase.** Plaintiff John Blase is a citizen and resident of Garland County, Arkansas. Plaintiff Blase authored and published *Touching Wonder: Recapturing the Awe*

of Christmas, and holds the copyright in his personal name (bearing Registration Number TX0007089143, registered on January 4, 2010).

25. **Defendant The EleutherAI Institute.** Defendant EleutherAI is a corporation formed in the State of Delaware and based in New York City.[9] EleutherAI is a self-described grassroots collective of natural language processing ("NLP") researchers who are explicitly oriented towards open sourcing the datasets related to the building of LLMs. Upon information and belief, EleutherAI has also specifically subjected itself to jurisdiction and venue in the State of New York in connection with providing AI services such as Books3 and the Pile.

26. **Defendant Bloomberg, L.P.** Defendant Bloomberg, L.P. is a limited partnership formed in the State of Delaware, with its principal place of business at Bloomberg Tower, 731 Lexington Ave., New York, NY 10022. Bloomberg, L.P. is a privately-held financial, software, data and media company which provides a raft of media services and related products, including TV, internet and radio news programing, financial terminal services, and other internet-based political, regulatory, and financial services.

27. **Defendant Bloomberg Finance, L.P.** Defendant Bloomberg Finance, L.P. is a limited partnership formed in the State of Delaware, with its principal place of business at 731 Lexington Ave, New York, New York 10022. Bloomberg Finance describes itself as "global business and financial information and news leader," giving "influential decision makers a critical edge by connecting them to a dynamic network of information, people and ideas."

28. **Defendant Meta, Inc.** Defendant Meta is a corporation formed in Delaware, with its principal place of business at 1601 Willow Rd Menlo Park, California 94025. Formerly known as Facebook, and now doing business as "Meta," Meta describes itself as a company dedicated to

[9] Castelvecchi, *supra* at 891.

giving "people the power to build community and bring the world closer together." Meta claims to "build technology that helps people connect and share, find communities, and grow businesses" and provides "useful and engaging products" to "enable people to connect and share with friends and family through mobile devices, personal computers, virtual reality headsets, and wearables."

29. **Defendant Microsoft Corporation.** Microsoft is a corporation formed in Delaware, with its principal place of business at One Microsoft Way, Redmond, Washington 98052. Microsoft describes itself as "a technology company whose mission is to empower every person and every organization on the planet to achieve more." Microsoft strives "to create local opportunity, growth, and impact in every country around the world" and claims to be "creating the platforms and tools, powered by artificial intelligence ('AI'), that deliver better, faster, and more effective solutions to support small and large business competitiveness, improve educational and health outcomes, grow public-sector efficiency, and empower human ingenuity."

FACTUAL ALLEGATIONS

A. Artificial Intelligence and the Training of Large Language Models

30. LLMs represent a significant advancement in the field of AI and have gained widespread attention for their impressive abilities to generate human-like text and perform various natural-language processing tasks. These models, powered by generative AI techniques, have revolutionized numerous industries and applications.

31. LLMs are AI systems designed to understand and generate human language. They are part of a broader category of generative artificial intelligence, which encompasses AI systems capable of creating content that closely resembles human-generated output. These models are typically built upon deep-learning architectures, with neural networks at their core.

32. Generative AI is a subset of artificial intelligence that focuses on creating AI systems capable of generating new data or content. Generative AI is fundamentally different from traditional AI, which often involves supervised learning, where models learn from labeled data to make predictions. Generative AI, on the other hand, involves training models to generate new content autonomously.

33. LLMs work by leveraging neural networks, specifically recurrent neural networks (RNNs) or more commonly, transformer-based architectures. LLMs rely upon interconnected steps to work:

a. LLMs are trained on vast datasets containing text from the internet, books, articles, and other sources. The training data is used to teach the model grammar, vocabulary, context, and various language patterns.

b. After collecting a training dataset, AI companies "tokenize," or break down the text contained in a training dataset into tiny pieces called "tokens." Tokens can be as short as a single letter or as long as a word or even a phrase, which helps the LLM understand the structure of language. Tokenization is important because it allows the LLM to view letters, words, and phrases as pieces of a puzzle, and to learn through iteration how these puzzle pieces fit together.

c. The LLM is then "trained" on this data: it is shown one sentence at a time, and it tries to guess what the next word or token should be. After it is given the correct solution, the LLM remembers both the training material it was provided and whether or not it produced the correct solution.

d. This process is followed hundreds of millions or even billions of times, and with each iteration, the LLM becomes slightly better at predicting the correct solution required for a certain input. Training an LLM is extremely data-intensive and requires a vast amount of easily-accessible plain text data to be effectively trained.

e. LLMs use deep neural networks with numerous layers and parameters. A "transformer architecture," which is a type of foundational neural network architecture with certain key components that allow LLMs to weigh the importance of different parts of output sequences, is used to process input and generate output.

f. Transformers incorporate a crucial component known as the "attention mechanism," which allows the model to focus on relevant parts of the input text when generating output. This mechanism enables the model to maintain context and coherence in longer pieces of text.

g. The model is trained to predict the next token in a sequence based on the context provided by the previous tokens. This process, called autoregression, helps the model learn to generate text that makes sense and flows naturally.

h. After pretraining on a massive dataset, LLMs can be fine-tuned for specific tasks, such as language translation, text generation, or question answering. Fine-tuning involves training the model on a smaller, task-specific dataset.

34. LLMs can generate coherent and contextually-relevant text, aiding in content creation, such as automated news articles, reports, and creative writing.

35. Corporations like the Defendants are in a rush to develop their own LLMs for several, compelling reasons.

a. First, being at the forefront of LLM technology provides a significant competitive edge. Companies that can leverage these models effectively can improve their products and services, leading to increased market share and customer loyalty.

b. LLMs can also automate a wide range of tasks, reducing labor costs and increasing efficiency. This is particularly valuable in industries like customer support, content generation, and data analysis.

c. Companies also deploy LLMs to create and enable hyper-personalized experiences, improving user engagement and satisfaction. “Recommender systems” powered by these models can provide tailored content and product recommendations, thus enhancing marketability and revenue generating opportunities.

d. LLMs also help companies analyze vast volumes of text data, extracting valuable insights and trends. This allows companies to make data-driven decisions and understand customer sentiment at a significantly reduced cost.

e. LLMs can handle large volumes of inquiries and requests simultaneously, allowing companies to scale their services without proportionally increasing their human workforce, which, again, saves the company money and drives revenues.

f. LLMs can also create content at scale, from news articles to product descriptions, saving time and resources for content-heavy industries.

36. Developing LLMs not only increases profits by allowing companies to make new and personalized offerings to their customers, but they also save companies money by reducing their reliance on a human workforce.

B. Pirate Troves Become Open-Source Training Materials

37. While LLMs can make and save corporations massive amounts of money, corporations are also incentivized to reduce the cost of creating and training LLMs.

38. Because the datasets necessary to effectively train LLMs are so vast, companies such as the Defendants have sought ways to constrain the cost of training their LLMs.

39. Enter "Books3," a plaintext dataset including the full text of approximately 197,000 non-fiction books and fictional novels published in the last twenty years, made available on the internet through a number of disparate sources.

40. Books3[10] was originally compiled by Shawn Presser, an independent developer, along with a team of "AI enthusiasts," in an effort "to give independent developers 'OpenAI-grade training data.'"[11] The name "Books3" is a reference to a paper published by LLM pioneer OpenAI in 2020 that mentioned two "internet-based books corpora" called Books1 and Books2, which OpenAI had used to train its own GPT-3 LLM.

41. Presser created Books3 in hopes that it would allow any developer to create generative-AI tools.

42. To create the dataset, Presser downloaded a copy of Bibliotik[12] from The-Eye.eu[13] and updated a program written more than a decade ago by hacker Aaron Swartz to convert the books from ePub format (a standard for ebooks) to plain text—a necessary change for the books to be used as training data.

43. After the creation of Books3, it was compiled into a larger dataset called "The Pile," which was then hosted on the internet by Defendant EleutherAI as a free, open-source data set for

[10] *See Dataset Card for the_pile_books3*, available at https://huggingface.co/datasets/the_pile_books3 ("This dataset contains all of bibliotik in plain .txt form, aka 197,000 books processed in exactly the same way as did for bookcorpusopen (a.k.a. books1). seems to be similar to OpenAI's mysterious 'books2' dataset referenced in their papers.").

[11] *See* Katie Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED, Sep. 4, 2023 (available at https://www.wired.com/story/battle-over-books3/).

[12] Bibliotik is, itself a "shadow library" of pirated, unlicensed copyright-protected works. *See* Claire Woodcock, *'Shadow Libraries' Are Moving Their Pirated Books to The Dark Web After Fed Crackdowns*, Vice (Nov. 30, 2022).

[13] This website's self-proclaimed purpose is to "suck up and serve large datasets." https://the-eye.eu/. It stopped hosting The Pile after receiving a DMCA takedown notice from a Danish anti-piracy company. Knibbs, n. 8 *supra*.

the training of LLMs. Alongside Books3, The Pile also includes many other unlicensed and copyrighted works, including YouTube-video subtitles, documents and transcriptions from the European Parliament, English Wikipedia, emails sent and received by Enron Corporation employees before its 2001 collapse, and more.

44. Researchers and journalists have dug into the Books3 data set and verified the presence of hundreds of thousands of books in that data set, including books authored by Plaintiffs and Class Members.[14]

45. Presser did not obtain licenses to use Plaintiffs' copyright-protected works before including them in the Books3 dataset and did not obtain any legal right (through assignment, license, or otherwise) to disseminate or otherwise distribute Plaintiffs' copyright-protected works.

46. In fact, Presser openly joked about the Digital Millennium Copyright Act, stating that he and his AI enthusiasts had "an interesting DMCA policy. I urge you to read it."[15] The link to the "policy" takes individuals to a loop of a profane performance that is 10 minutes and 50 seconds long. Presser's post was cited in the EleutherAI paper and was publicly available.

47. EleutherAI, despite knowing the contents of the Books3 dataset, incorporated it into The Pile and hosted this library of unlicensed content, including the copyright-protected works of the Plaintiffs and Class Members.

48. Indeed, EleutherAI continues to distribute Books3 through its website (https://pile.eleuther.ai/). The Books3 dataset is also available from a popular AI project hosting

[14] Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, The Atlantic (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/.

[15] *See* Shawn Presser, Oct. 25, 2020 Tweet, *available at* https://twitter.com/theshawwn/status/1320282154488766464. The link is to a profane performance and is, likely, a joke acknowledging the copyright violation.

service called "Hugging Face," a "New York City-based company that aims to expand access to AI" and "lists more than 100 open-source LLMs on its website" (https://huggingface.co/datasets/the_pile_books3).[16]

C. LLaMa: Microsoft and Meta Combine Forces

49. Prior to the amalgamation and dissemination of Books3, Defendant Meta was hard at work on the creation and training of its own LLM.

50. Meta's large language model tool Large Language Model Meta AI (or "LLaMa") is a family of LLMs, the release of which was announced by Meta in February of 2023 via a blog post and a paper describing the model's training, architecture, and performance.[17]

51. Meta's first version of LLaMa included four models, which were trained with 7, 13, 33 and 65 billion parameters.

52. Meta acknowledged that its original dataset for LLaMa came from the Pile and described Books3 as a "publicly available dataset for training large language models."[18] Meta failed to acknowledge that the "publicly available dataset" contained copyrighted works of authorship that were pirated by the creators and custodians of Books3, including EleutherAI.

53. Using data from the Pile would enable Meta to create its LLM faster and more efficiently, and save the company a substantial amount of time, resources, and money because it would not have to pay for source material, including copyrighted source material by the Plaintiffs and Class Members.

[16] Castelvecchi, *supra* at 891.

[17] Research, *Introducing LLaMA: A foundational, 65-billion-parameter large language model*, Meta (Feb. 24, 2023), https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

[18] Hugo Touvron, *et al.*, *LLaMA: Open and Efficient Foundation Language Models* (Feb. 2023), *available at* https://browse.arxiv.org/pdf/2302.13971.pdf, at 2 ("We include. . . the Books3 section of ThePile. . . a publicly available dataset for training large language models.").

54. The original release of LLaMa was powerful: its developers reported that the thirteen-billion-parameter model's performance on most NLP benchmarks exceeded that of the much larger OpenAI's GPT-3 (with 175B parameters) and that the largest model was competitive with state-of-the-art models such as PaLM and Chinchilla.[19]

55. In March 2023, the LLaMA language models were leaked to a public internet site and have continued to circulate. Meta has not disclosed what role it had, if any, in the leak, or how and why the leak occurred.

56. Meta issued a "DMCA takedown notice" to a programmer on GitHub who had released a tool that helped users download the leaked LLaMA language models. In the notice, Meta asserted copyright over the LLaMA language models.[20]

57. After the initial success of Meta's LLaMa, Meta announced in July 2023 that it was partnering with Defendant Microsoft to develop LLaMa 2, the next iteration of the LLM, explicitly designed to compete with OpenAI's ChatGPT.[21]

58. In August 2023, Meta released LLaMa 2, a new version of the model with several improvements, which it stated that it would make available for commercial use.

59. Meta claimed in its initial press releases that LLaMA 2 model had been "red-teamed," or tested for safety by "generating adversarial prompts to facilitate model fine-tuning," both internally and externally.[22]

[19] Touvron, *supra* at 11.

[20] GitHub, dmca/2023/03/2023-03-21-meta.md, *available at* https://github.com/github/dmca/blob/master/2023/03/2023-03-21-meta.md.

[21] John Montgomery, *Microsoft and Meta expand their AI partnership with Llama 2 on Azure and Windows*, Microsoft Official Blog (July 18, 2023), https://blogs.microsoft.com/blog/2023/07/18/microsoft-and-meta-expand-their-ai-partnership-with-llama-2-on-azure-and-windows/.

[22] *Meta and Microsoft Introduce the Next Generation of Llama*, Meta (July 18, 2023), https://ai.meta.com/blog/llama-2/.

60. Meta also disclosed how the models are evaluated and tweaked. Despite this, there is no indication in Meta's press releases, papers, or elsewhere that the Pile dataset, and especially Books3, was excluded from LLaMa 2's training data.

61. Microsoft and Meta benefitted greatly from prior iterations of the LLM using Books3 and information from the Pile because they did not have to spend additional time, money, and resources to train an LLM from scratch with their own content or properly-licensed content. Instead, they were able to incorporate sophisticated datasets, which included the pirated copyright-protected materials in Books3, as part of the LLM's training process, without having to compensate the authors.

62. Upon information and belief, LLaMa 2 continued to utilize Plaintiffs' copyright-protected works in the Books3 training dataset in the same fashion that the initial LLaMa models did.

D. BloombergGPT: The First Finance-Oriented LLM

63. In the aftermath of the release of the enormously popular OpenAI GPT-3, Bloomberg also began work on its own LLM, which it announced on March 30, 2023.

64. Bloomberg introduced BloombergGPT, the world's first LLM "built from scratch for finance."[23]

65. Bloomberg released a research paper detailing how the 50-billion-parameter LLM had been trained specifically on a wide range of financial data "to support a diverse set of natural language processing (NLP) tasks within the financial industry."[24]

[23] *Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance*, Bloomberg Professional Services (Mar. 30, 2023), https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/.

[24] *Id.*

66. Bloomberg boasted that it had been "a trailblazer in its application of AI, Machine Learning, and NLP in finance."[25]

67. Comparing its LLM's performance to other LLMs, including OpenAI's GPT-3, Bloomberg touted its performance, and credited its "large training corpus with over 700 billion tokens."[26]

68. Part of that training corpus was Books3, which Bloomberg used to assist its LLM in learning how to recognize, parse, and respond in natural language.[27]

69. Using data from Books3 would enable Bloomberg to create its LLM faster and more efficiently, and save the company a substantial amount of time, resources, and money because it would not have to pay for source material.

70. Indeed, shortly after its announcement, a spokesperson for Bloomberg confirmed via email that Books3 was used to train the initial model of BloombergGPT and added, "We will not include the Books3 dataset among the data sources used to train future versions of BloombergGPT."

71. But Bloomberg knows that its representation is hollow; LLM training is iterative and builds on prior versions. Because the Books3 training dataset has already been used to train prior versions of the BloombergGPT.

72. Plaintiffs' copyright-protected works have been baked into Bloomberg's LLM, and all subsequent versions: it cannot simply weed out the benefit it has illegally gained from scanning the text of the copyright-protected works written by Plaintiffs and Members of the Class.

[25] *Id.*

[26] *Id.*

[27] Andrew Lukyanenko, *Paper Review: BloombergGPT: A Large Language Model for Finance*, Medium, https://betterprogramming.pub/paper-review-bloomberggpt-a-large-language-model-for-finance-39d771efdedc (last accessed Oct. 6, 2023).

73. Bloomberg benefitted substantially from its theft. Bloomberg did not compensate Plaintiffs for their copyrighted material while creating a lucrative product.

74. Without starting over from a previous version which did not use a training dataset including Books3, Bloomberg's LLM is tainted with illicitly obtained, copyright-protected material.

E. The Cost to the Plaintiffs

75. AI driven LLMs have become one of the fastest-emerging technologies on the internet. For instance, OpenAI is expected to generate over $1 billion in revenue in the next 12 months.[28] Other research institutions believe the AI industry is expected to be as large as $1.8 trillion by 2030.[29]

76. Companies see immense value in the speed with which LLMs can research vast troves of research and produce written product—there are very few industries that have not begun at least informally adopting the use of AI-driven LLMs to boost productivity and performance.

77. Companies in the industry have acknowledged the value as well; to say nothing of the partnership between defendants Microsoft and Meta to continue to develop the LLaMa LLM, Microsoft itself has already invested approximately $13 billion in ChatGPT's creator, OpenAI.

[28] Amir Efrati and Aaron Holmes, *OpenAI Passes $1 Billion Revenue Pace as Big Companies Boost AI Spending*, The Information (Aug. 29, 2023), https://www.theinformation.com/articles/openai-passes-1-billion-revenue-pace-as-big-companies-boost-ai-spending.

[29] "Artificial Intelligence Market to Hit $1,811.75 Billion by 2030 Grand View Research, Inc." PR Newswire, *Bloomberg* (July 3, 2023), https://www.bloomberg.com/press-releases/2023-07-03/artificial-intelligence-market-to-hit-1-811-75-billion-by-2030-grand-view-research-inc#:~:text=of%20Brazil%20Indictment-,Artificial%20Intelligence%20Market%20to%20Hit%20%241%2C811.75%20Billion,%3A%20Grand%20View%20Research%2C%20Inc.

78. Defendants have illicitly gained an enormous amount of value from their unauthorized use of Plaintiffs' copyright-protected works—without use of the Books3 data set, Defendants would not have been able to train their LLMs to recognize and respond to queries in a fashion that is useful to average users.

79. More importantly, the Books3 dataset, and Plaintiffs' copyright-protected works, now serve as a baseline for all future LLM models created by Defendants—Defendants cannot simply agree to strip the Plaintiffs' copyright-protected works out of any future data sets, because Defendants will continue to benefit from unlawfully acquiring and using Plaintiffs' copyright-protected works.

80. Plaintiffs have not been compensated for any of this. Their books were pirated, converted to plain text, and used as a way to train LLMs without providing them with the value for their work or licensing fees.

81. Defendants have infringed upon Plaintiffs' copyrights and have profited enormously from the use of their copyright-protected works.

CLASS ALLEGATIONS

82. Plaintiffs bring this action pursuant to the provisions of Rules 23(a), 23(b)(2), and 23(b)(3) of the Federal Rules of Civil Procedure, on behalf of themselves and the following proposed Class:

> All persons or entities in the United States that own a United States copyright in any work that was used as training data for the Defendants LLMs during the from October, 2020 to the present.

83. Excluded from the Class are Defendants, its employees, officers, directors, legal representatives, heirs, successors, wholly- or partly-owned, and its subsidiaries and affiliates; proposed Class counsel and their employees; the judicial officers and associated court staff

assigned to this case and their immediate family members; all persons who make a timely election to be excluded from the Class; governmental entities; and the judge to whom this case is assigned and his/her immediate family.

84. This action has been brought and may be properly maintained on behalf of the Class proposed herein under Federal Rule of Civil Procedure 23.

85. **Numerosity**. Federal Rule of Civil Procedure 23(a)(1): The members of the Class are so numerous and geographically dispersed that individual joinder of all Class members is impracticable. On information and belief, there are at least tens of thousands of members in the Class. The Class members may be easily derived from Defendants' records and other publicly-available information about the copyrighted works included in the Books3 dataset.

86. **Commonality and Predominance**. Federal Rule of Civil Procedure 23(a)(2) and 23(b)(3): This action involves common questions of law and fact, which predominate over any questions affecting individual Class members, including, without limitation:

a. Whether Defendants violated the copyrights of Plaintiffs and the Class when they downloaded copies of Plaintiffs' and the Class's Infringed Works and used them to train LLMs;

b. Whether the LLMs are themselves infringing derivative works based on Plaintiffs' and the Class's Infringed Works;

c. Whether the text outputs of the LLMs are infringing derivative works based on Plaintiffs' Infringed Works;

d. Whether Defendants violated the DMCA by removing copyright-management information from Plaintiffs' and the Class's Infringed Works;

e. Whether Defendants were unjustly enriched by the unlawful conduct alleged herein;

f. Whether the Defendants conduct constitutes a conversion under relevant law.

g. Whether Defendants' conduct alleged herein constitutes common unfair competition;

h. Whether any affirmative defense excuses Defendants' conduct;

i. Whether any statutes of limitation limits Plaintiffs' and the Class's potential for recovery;

j. Whether Plaintiffs and the other Class members are entitled to equitable relief, including, but not limited to, restitution or injunctive relief; and

k. Whether Plaintiffs and the other Class members are entitled to damages and other monetary relief and, if so, in what amount.

87. **Typicality**. Federal Rule of Civil Procedure 23(a)(3): Plaintiffs' claims are typical of the other Class members' claims because, among other things, all Class members were comparably injured through Defendants' wrongful conduct as described above.

88. **Adequacy**. Federal Rule of Civil Procedure 23(a)(4): Plaintiffs will fairly and adequately represent and protect the interests of Class Members, including those from states and jurisdictions where they do not reside. Counsel representing the Plaintiffs in this action are competent and experienced in litigating class actions and have been appointed lead counsel by many different courts in many other class action suits.

89. **Declaratory and Injunctive Relief**. Federal Rule of Civil Procedure 23(b)(2): Defendants have acted or refused to act on grounds generally applicable to Plaintiffs and the other members of the Class, thereby making appropriate final injunctive relief and declaratory relief with

respect to the Class as a whole.

90. **Superiority**. Federal Rule of Civil Procedure 23(b)(3): A class action is superior to any other available means for the fair and efficient adjudication of this controversy, and no unusual difficulties are likely to be encountered in the management of this class action. The damages or other financial detriment suffered by Plaintiffs and the other Class members are relatively small compared to the burden and expense that would be required to individually litigate their claims against Defendants, so it would be impracticable for the members of the Class to individually seek redress for Defendants' wrongful conduct. Even if Class members could afford individual litigation, the court system could not. Individualized litigation creates a potential for inconsistent or contradictory judgments and increases the delay and expense to all parties and the court system. By contrast, the class action device presents far fewer management difficulties, and provides the benefits of single adjudication, economy of scale, and comprehensive supervision by a single court.

CAUSES OF ACTION

FIRST CAUSE OF ACTION
DIRECT COPYRIGHT INFRINGEMENT, 17 U.S.C. § 106, *et seq*.

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

91. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

92. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class against Defendants.

93. As the owners of the registered copyrights in the Infringed Works, Plaintiffs and the Class hold the exclusive rights to those works under the Copyright Act, 17 U.S.C. § 106.

94. To train Defendants' LLMs, Defendants created, copied, maintained and/or utilized the Books3 dataset, which includes unlicensed copies of the Infringed Works obtained from shadow libraries.

95. Plaintiffs and Class members never authorized Defendants to make copies of their Infringed Works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works).

96. At all times during the Class Period, Plaintiffs and members of the Class maintained all rights under copyright law.

97. Defendants used the Infringed Works without Plaintiffs' permission and without the permission of the Class members.

98. Because the LLMs cannot function without the expressive information extracted from the Infringed Works and are retained inside the Books3 database and the LLMs, these datasets and models are themselves infringing derivative works, made without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

99. Plaintiffs and the members of the Class have been injured by Defendants' acts of direct copyright infringement. Plaintiffs and the Class are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

SECOND CAUSE OF ACTION
VICARIOUS COPYRIGHT INFRINGEMENT, 17 U.S.C. § 106

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

100. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

101. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class against Defendants.

102. Because the output of the LLMs is based on expressive information extracted from Plaintiffs' and the Class's Infringed Works, every output of the LLMs is an infringing derivative work, made without permission and in violation of their exclusive rights under the Copyright Act.

103. Defendants have the right and ability to control the output of the LLMs and or the dataset contained in Books3. Defendants have benefited financially from the infringing output of the language models. Therefore, every output from the language models constitutes an act of vicarious copyright infringement for which all Defendants may be held liable.

104. Plaintiffs and the Class have been injured by Defendants' acts of vicarious copyright infringement. Plaintiffs and the Class are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

THIRD CAUSE OF ACTION
DIGITAL MILLENNIUM COPYRIGHT ACT – REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION, 17 U.S.C. § 1202(B)

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

105. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

106. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class against Defendants.

107. Plaintiffs included one or more forms of copyright-management information ("CMI") in each of the Infringed Works, including: copyright notice, title and other identifying information, or the name or other identifying information about the owners of each book, terms and conditions of use, and identifying numbers or symbols referring to CMI.

108. Without the authority of Plaintiffs and the Class, Defendants copied the Infringed Works and used them as training data large language models, including the LLMs. By design, the training process does not preserve any CMI. Therefore, Defendants intentionally removed CMI

from the Infringed Works in violation of 17 U.S.C. § 1202(b)(1).

109. Without the authority of Plaintiffs and the Class, Defendants created derivative works based on the Infringed Works. By distributing these works without their CMI, Defendants violated 17 U.S.C. § 1202(b)(3).

110. By falsely claiming that they have sole copyright in the language models—which they cannot, because the language models are infringing derivative works—Defendants violated 17 U.S.C. § 1202(a)(1).

111. Defendants knew or had reasonable grounds to know that this removal of CMI would facilitate copyright infringement by concealing the fact that every output from their language models is an infringing derivative work, synthesized entirely from expressive information found in the training data.

112. Plaintiffs and the Class have been injured by Defendants' removal of CMI. Plaintiffs and the Class are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

FOURTH CAUSE OF ACTION
CONVERSION

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

113. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

114. Each of the Defendants intentionally exercised unauthorized dominion or control over the property of the Plaintiffs and the Class by way of maintaining, utilizing, or otherwise using the Infringed Works by and through the Books3 dataset.

115. The intentional acts of Defendants were taken with the intent to deprive Plaintiffs and the Class of their rights.

116. The intentional acts of the Defendants constitute a conversion for which the Plaintiffs and the Defendants have suffered damage. The Defendants' conversion is ongoing.

117. The intentional conversion of the Plaintiffs and the Class's property by the Defendants was done with sufficient malice and forethought such that punitive damages must be awarded.

FIFTH CAUSE OF ACTION
NEGLIGENCE

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

118. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

119. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class against Defendants.

120. Defendants owed a duty of care toward Plaintiffs and the Class based upon Defendants' relationship to them. This duty is based upon Defendants' obligations, custom and practice, right to control information in its possession, exercise of control over the information in its possession, authority to control the information in its possession, and the commission of affirmative acts that result in said harms and losses.

121. Defendants breached its duties by negligently, carelessly, and recklessly collecting, maintaining and controlling Plaintiffs' and Class members' Infringed Works and engineering, designing, maintaining and controlling systems, which are trained on Plaintiffs' and the Class's Infringed Works without their authorization.

122. Defendants owed Plaintiffs and the Class a duty of care to maintain the Infringed Works once collected and ingested for training the Defendants' LLMs.

123. Defendants also owed Plaintiffs and the Class members a duty of care to not use the Infringed Works in a way that would foreseeably cause Plaintiffs and the Class members injury, for instance, by using the Infringed Works to train the Defendants' LLMs.

124. Defendants breached their duties by maintaining the Infringed Works and otherwise using them to train LLMs.

125. This breach of duties has damaged Plaintiffs and the Class in amounts to be proven at trial.

SIXTH CAUSE OF ACTION
UNJUST ENRICHMENT

Brought on behalf of Plaintiffs, individually and on behalf of the Class, against all Defendants

126. Plaintiffs hereby incorporate by reference the allegations contained in paragraphs 1-90 of this Complaint.

127. Plaintiffs and the Class have invested substantial time and energy in creating the Infringed Works.

128. Defendants have unjustly utilized access to the Infringed Materials to train LLMs, including those maintained by the Defendants.

129. Plaintiffs did not consent to the unauthorized use of the Infringed Materials to train the LLMs or any other artificial intelligence system.

130. By using Plaintiffs' Infringed Works to train the LLMs, Plaintiffs and the Class were deprived of the benefits of their work, including monetary damages.

131. Defendants derived or intend to derive profit and other benefits from the use of the Infringed Materials to train the LLMs.

132. It would be unjust for Defendants to retain those benefits.

133. The conduct of Defendants is causing and, unless enjoined and restrained by this Court, will continue to cause Plaintiffs and the Class great and irreparable injury that cannot fully be compensated or measured in money.

REQUEST FOR RELIEF

WHEREFORE, Plaintiffs, individually and on behalf of members of the Class defined above, respectfully request that the Court enter judgment against Defendants and award the following relief:

a. Certification of this action as a class action pursuant to Rule 23 of the Federal Rules of Civil Procedure, declaring Plaintiffs as the representatives of the Class, and Plaintiffs' counsel as counsel for the Class;

b. An order awarding declaratory relief and temporarily and permanently enjoining Defendants from continuing the unlawful and unfair business practices alleged in this Complaint and to ensure that all applicable information set forth in 17 U.S.C. § 1203(b)(1) is included when appropriate;

c. An award of statutory and other damages under 17 U.S.C. § 504 for violations of the copyrights of Plaintiffs and the Class by Defendants;

d. An award of statutory damages under 17 U.S.C. § 1203(b)(3) and 17 U.S.C. § 1203(c)(3), or in the alternative, an award of actual damages and any additional profits under 17 U.S.C. § 1203(c)(2);

e. A declaration that Defendants are financially responsible for all Class notice and the administration of Class relief;

f. An order awarding any applicable statutory and civil penalties;

g. An order requiring Defendants to pay both pre- and post-judgment interest on any amounts awarded;

h. An award of punitive damages;

i. An award of costs, expenses, and attorneys' fees as permitted by law; and

j. Such other or further relief as the Court may deem appropriate, just, and equitable.

JURY DEMAND

Plaintiffs and the Class demand a trial by jury on all issues so triable.

Date: October 17, 2023

Respectfully Submitted,

By: /s/ Greg G. Gutzler
Greg G. Gutzler
DICELLO LEVITT LLP
485 Lexington Avenue, Tenth Floor
New York, New York 10017
Telephone: (646) 933-1000
Facsimile: (646) 494-9648
ggutzler@dicellolevitt.com

Adam J. Levitt
Amy E. Keller*
James A. Ulwick*
DICELLO LEVITT LLP
Ten North Dearborn Street, Sixth Floor
Chicago, Illinois 60602
Tel. (312) 214-7900
alevitt@dicellolevitt.com
akeller@dicellolevitt.com
julwick@dicellolevitt.com

Mr. Seth Haines*
Mr. Timothy Hutchinson*
Ms. Lisa Geary*
RMP, LLP
5519 Hackett Street, Suite 300
Springdale, Arkansas 72762
Tel: (479) 443-2705
shaines@rmp.law
thutchinson@rmp.law
lgeary@rmp.law

Mr. Scott Poynter*
POYNTER LAW GROUP
407 President Clinton Avenue, Suite 201
Little Rock, Arkansas 72201
Tel: (501) 812-3943
scott@poynterlawgroup.com

Counsel for Plaintiffs and the Proposed Class

* *Pro hac vice* applications forthcoming